Klasifikasi Daerah Tertinggal di Indonesia Menggunakan Algoritma SVM dan *k*-NN

Classification of Underdeveloped Areas in Indonesia Using the SVM and k-NN Algorithms

Harun Al Azies*), Gangga Anuraga
Department of Statistics, Faculty of Mathematics and Natural Sciences, PGRI Adi Buana University
*E-mail: harunalazies@gmail.com

ABSTRACT

The determination or classification of underdeveloped areas essentially consists of classifying several observations taking into account existing indicators. The classification method used is K-Nearest Neighbor (*k*-NN) and Support Vector Machines (SVM). This study aims to analyze the accuracy of the classification between SVM and *k*-NN algorithms in the classification of underdeveloped areas in Indonesia. The data source used in this study is secondary data obtained from the Central Bureau of Statistics (BPS). The data used are 514 districs and municipalities of Indonesia. After analysis, the conclusion is that there are 122 districs and municipalities that are left behind out of a total of 514 districs and municipalities in Indonesia. The most underdeveloped areas are on the island of Papua, followed by the areas of the islands of Bali and Nusa Tenggara, and Sulawesi. Based on the results of the classification of underdeveloped areas using the method SVM with the kernel RBF has the best results with the parameters C = 1 and γ = 0.05 while the results of the classification of underdeveloped areas using the method *k*-NN obtains the best results with k = 15 Based on the results of classification of underdeveloped areas using the SVM and the *k*-NN method, including the level of classification is very good. The two methods compared have the same precision value of 92.2% and can be used to determine the classification of underdeveloped areas.

Keywords: classification, machine learning, supervised learning, underdeveloped areas.

PENDAHULUAN

Kesenjangan pembangunan dan perkembangan antara wilayah masih terjadi di Indonesia, sehingga masih terdapat wilayah-wilayah yang sudah maju dan berkembang pesat, namun berbanding terbalik dengan wilayah-wilayah yang masih kurang berkembang dan bahkan termasuk kedalam wilayah tertinggal. Penentuan atau pengklasifikasian kabupaten tertinggal dan tidak tertinggal adalah metode mengelompokkan wilayah berdasarkan indikator yang telah ditetapkan (Purwandari, 2017). Permasalahan ini dapat diselesaikan menggunakan salah satu metode dalam *machine learning.*

Machine Learning (ML) atau pembelajaran mesin merupakan pendekatan dalam *Artificial Intelligence* (AI) (Russell, 2016). ML menjadi salah satu bidang ilmu komputer yang tumbuh paling cepat, dengan aplikasi yang luas jangkauannya (Shalev-Shwartz, 2014). Algoritma *machine learning* memiliki beberapa jenis diantaranya *supervised learning algorithms, unsupervised learning algorithms, semi-supervised learning* dan *reinforcement learning* (Smola, 2008). Penelitian ini berfokus pada salah satu algoritma *machine learning* yaitu *supervised learning*. *Supervised learning* adalah algoritma khusus untuk klasifikasi yang (Kotsiantis, 2007), cara kerjanya memetakan input ke sebuah *output* yang diinginkan atau algoritma (Ayodele, 2010).

Algoritma dalam *supervised machine learning* yang menjadi fokus penelitian ini adalah *Support Vector Machines* (SVM) dan *k-Nearset Neighbor* (*k*-NN). *Support vector machine* adalah metode yang dikenalkan oleh Vapnik pada tahun 1992 (Gunn, 1998) dengan cara kerja dasarnya adalah mengklasifikasikan suatu kasus dengan memaksimalkan batas-batas *hyperplane* (Abe, 2010). Pemilihan algoritma SVM adalah terkait performanya dalam mengklasifikasikan suatu *pattern*/pola, selain itu kelebihan algoritmaini mencegah terjadinya permasalahan dimensionalitas (Tan *et al*., 2019). Sedangkan metode klasifikasi algoritma *k*-NN dalam penelitian yang dilakukan oleh (Deng *et al*., 2016) algoritma *k*-NN digunakan untuk mengklasifikasikan setiap sampel pengujian berdasarkan *k* tetangga

terdekat di cluster data terdekat. Kluster yang pusatnya memiliki jarak Euclidean minimum dari sampel uji adalah yang paling dekat. Konsep jarak Euclidean ini memperlakukan semua variable adalah bebas (tidak berkorelasi) (James *et al*., 2013).

Penelitian mengenai *Machine Learning* dengan membandingkan algoritma klasifikasi semakin banyak dilakukan. (Delgado *et al*., 2014) melakukan penelitian dengan menggunakan 179 jenis klasifier yang diterapkan pada 121 kumpulan data dari basis data *UCI Machine Learning Repository* hasil evaluasi menunjukkan bahwa yang terbaik adalah metode *Random Forest* (RF) dan SVM dengan kernel Gaussian yang. (Guo *et al*., 2003) dengan menggunakan dataset dari *UCI Machine Learning Repository* melakukan penelitian tentang pendekatan berbasis model *k*-NN, hasil percobaan menunjukkan bahwa model berbasis k-NN merupakan metode yang cukup kompetitif untuk klasifikasidapat dibandingkan dengan C5.0 dan *k*-NN standard dengan hal akurasi klasifikasi yang baik, tetapi lebih efisien daripada *k*-NN standar. Selain itu (Jung *et al*., 2018) juga melakukan evaluasi kinerja dari tiga pengklasifikasi yaitu SVM, *distance-weighted k-nearest neighbour* (WKNN), dan *decision tree* (DT) dengan menggunakan data dari solusi set sensor yang dioptimalkan dan tidak dioptimalkan.

Pada permasalahan di Indonesia (Fernanda *et al*., 2019) melakukan perbandingan metode klasifikasi pada permasalahan hipertensi. Metode yang digunakan untuk menganalisis faktor risiko yang signifikan adalah regresi logistik dan *Classification and Regression Tree* (CART) dengan menggunakan metode yang sama yaitu regresi logistic Al Azies (2017) mendapatkan akurasi 95% untuk meng-klasifikasikan perilaku hidup bersih dan sehat (PHBS) Rumah Tangga Penderita TB di Wilayah Pesisir Kota Surabaya. (Puspitasari, 2018) Menerapkan SVM dan *k*-NN menggunakan SVR sebagai fitur seleksi pada analisis saham untuk Bursa Efek Indonesia. Demikian pula pada penelitian (Al Azies *et al*., 2019) melakukan penelitian untuk membandingkan kernel pada SVM dalam klasifikasi Indeks Pembangunan Manusia, hasil klasifikasi menunjukkan bahwa kernel *Radial Basis Function* (RBF) adalah metode yang sesuai untuk mengklasifikasikan IPM.

METODE

Sumber Data

Sumber data pada penelitian ini didasarkan dari data histori yang merupakan sumber data sekunder berupa 16 variabel penelitianyang didapatkan berdasar Perpres Nomor 131 Tahun 2015 dan yang digunakan digunakan oleh Kementerian Negara Pembangunan Daerah Tertinggal dan Transmigrasi sebagai indikator penetapan ketertinggalan daerah, data tersebut diperoleh dari Publikasi Badan Pusat Statistik.

Obyek dan Variabel Penelitian

Obyek pengamatan pada penelitian ini adalah kabupaten dan kota di Indonesia sebanyak 514. Variabel terbagi menjadi satu variabel respon dan variabel prediktor. Variabel prediktor yang digunakan sebanyak 15 variabel yang terbagi menjadi 6 kriteria. Adapaun variabel yang digunakan dijelaskan pada Tabel 1.

Tabel 1. Variabel Penelitian

Variabel	Skala
Y = 1 (Kabupaten/Kota Tertinggal) Y = 0 (Kabupaten/Kota Tidak Tertinggal)	Nominal
Indeks Kedalaman Kemiskinan (X_1)	Rasio
Indeks Keparahan Kemiskinan (X_2)	Rasio
Tingkat Pengangguran Terbuka (TPT) (X_3)	Rasio
Harapan Lama Sekolah (X_4)	Rasio
Usia Harapan Hidup (X_5)	Rasio
Banyaknya Desa/Kelurahan Menurut Ketersediaan Sistem Keuangan Desa (X_6)	Rasio
Banyaknya Desa/Kelurahan Menurut Keberadaan Sarana Kesehatan Rumah Sakit(X_7)	Rasio
Banyaknya Desa/Kelurahan Yang Mempunyai Sekolah Jenjang SMA/SMK (X_8)	Rasio
Persentase Rumah Tangga Menurut Sumber Air Minum Leding (X_9)	Rasio
Persentase Rumah Tangga Menurut Dan Sumber Penerangan PLN (X_{10})	Rasio
Rata-Rata Jarak Kabupaten/Kota ke Ibukota Provinsi (X_{11})	Rasio
Banyaknya Desa/Kelurahan Menurut Ketersediaan Angkutan Umum (X_{12})	Rasio
Jumlah Desa Yang Mengalami Banjir (X_{13})	Rasio
Jumlah Desa Yang Mengalami Gempa Bumi (X_{14})	Rasio
Jumlah Desa Yang Mengalami Tanah Longsor (X_{15})	Rasio

Langkah-langkah Penelitian

Tahapan analisis yang digunakandalam penelitian ini adalah sebagai berikut.

1. Menggunakan statistika deskriptif untuk melakukan eksplorasi data sebagai tujuan mengetahui gambaran umum kondisi persebaran ketertinggalan daerah di Indonesia.
2. Membagi data menjadi data training dan data testing. Dataset yang diperoleh dibagi menjadi 2 bagian yaitu *data training* (75%) dan *data testing* (25%). Dengan rincian pada Tabel 2.

Tabel 2. Komposisi Pembagian Dataset

	Daerah Tertinggal	Daerah Tidak Tertinggal	Total
Data Training	96	289	385 (75%)
Data Testing	26	103	129 (25%)

3. Mengklasifikasi status ketertinggalan wilayah dengan menggunakan algoritma SVM
   a. Melakukan optimasi parameter pada SVM untuk setiap jenis kernel (kernel linier, RBF dan Polinomial)
   b. Menyusun *confusion matrix*.
   c. Menghitung nilai akurasi untuk mengukur performa model.
4. Mengklasifikasi status wilayah tertinggal dengan menggunakan algoritma *k*-NN
   a. Melakukan optimasi parameter *k*
   b. Menghitung kuadrat jarak *euclid*(*query instance*) masing-masing objek terhadap training data yang diberikan, selanjutnya data diurutkan berdasarkan *euclidean distance* terkecil ke terbesar.
   c. Pemeringkatan hasil pengurutan sesuai dengan nilai k, lalu tentukan
   d. Pengkategorian atau pelabelan dari data yang telah diperingkatkan tersebut berdasarkan kategori tetangga terdekat yang paling banyak
5. Melakukan pemilihan algoritma terbaik berdasarkan performa klasifikasi.

HASIL DAN PEMBAHASAN

Gambaran Umum Ketertinggalan Daerah di Indonesia

Pemerintah setiap lima tahun sekali didalam RPJMN mengeluarkan status terbaru kondisi ketertinggalan wilayah di Indonesia. Indonesia terdiri dari 34 provinsi dan terbagi menjadi 514 kabupaten dan kota. Berdasarkan Peraturan Presiden yang dikeluarkan pada tahun 2015 Nomor 131 tentang penetapan status kabupaten dan kota tertinggal, dari 514 kabupaten dan kota di Indonesia 24 persen atau masih terdapat 122 wilayah yang masuk kedalam kategori wilayah tertinggal. Hasil pemetaan pada Gambar 1 menjelaskan persebaran wilayah tertinggal di Indonesia. Mayoritas kabupaten dan kota tertinggal di Indonesia (dilambangkan dengan warna kuning) didominasi merupakan kabupaten dan kota yang berada di wilayah timur Indonesia.

Gambar 1**.** Sebaran Ketertinggalan Daerah

Terdapat 103 kabupaten dan kota di wilayah timur Indonesia berstatus sebagai wilayah tertinggal, sementara itu 19 kabupaten dan kota lain merupakan wilayah berstatus tertinggal berada dikawasan Indonesia bagian barat. Berdasarkan Perpres Nomor 131 tahun 2015 juga dapat diketahui bahwa Provinsi Papua menjadi wilayah dengan jumlah kabupaten dan kota berstatus tertinggal terbanyak di Indonesia, terdapat 26 dari 29 kabupaten dan kota di Provinsi Papua berstatus sebagai wilayah tertinggal

Klasifikasi Daerah Tertinggal Menggunakan Algoritma *Support Vector Machine* (SVM)

Support Vector Machine (SVM) merupakan salah satu jenis dari *supervised machine learning* yang akan menjadi salah satu algoritma untuk mengklasifikasikan status ketertinggalan wilayah di Indonesia. Unit observasi pada penelitian ini adalah seluruh kabupaten dan kota di Indonesia yang berjumlah 514. Seperti dijelaskan pada langkah penelitian poin kedua, bahwa data kabupaten dan kota yang berjumlah 514 akan dibagi untuk dilakukan pemisahan menjadi data training sebanyak 75% dari keseluruhan 514 kabupaten dan kota, sementara itu sisa data akan digunakan sebagai data testing. Selanjutnya seperti dijelaskan pada langkah penelitian poin ketiga untuk klasifikasi menggunakan algoritma SVM ini akan menggunakan tiga fungsi untuk mendapatkan algoritma terbaik berdasarkan akurasi ketepatan klasifikasinya. Tiga fungsi SVM yang digunakan pada penelitian ini adalah linier, RBF dan polynomial.

Klasifikasi Daerah Tertinggal Menggunakan Linear Kernel SVM

Klasifikasi pertama yaitu klasifikasi menggunakan kernel linier. Kernel linier adalah salah satu fungsi dalam SVM yang digunakan untuk karakteristik data yang terindikasi terklasifikasi secara linier. Setiap kernel dalam SVM memiliki perbedaan, perbedaan mendasar setiap kernel adalah pada parameter yang digunakan. Kernel linier pada SVM memiliki parameter C atau *Cost*. Kernel linier bekerja dengan cara mengoptimasi parameter Cuntuk mendapatkan akurasi klasifikasi terbaik dengan cara melakukan berbagai kombinasi model atau *trial and error*. Penentuan parameter terbaik SVM menggunakan linear kernel dapat dievaluasi melalui ukuran performa klasifikasi yang dapat diukur menggunakan akurasi. Hasil optimasi parameter C menggunakan fungsi kernel linier adalah sebagai berikut.

Tabel 3. Hasil Optimasi Pemilihan Parameter Terbaik Kernel Linier

Parameter (C)	Akurasi
10^{-4}	0.798
10^{-3}	0.814
10^{-2}	**0.915***
10^{-1}	0,891
1	0.899

Ket : *) Parameter terpilih dengan nilai akurasi terbesar

Tabel 3 menunjukkan hasil optimasi parameter C menggunakan kernel linier, hasil tersebut didapat menggunakan data training yaitu 75 persen dari dataset.Berdasarkan hasil optimasi tersebut didapatkan nilai akurasi terbaik yaitu 0.915 atau 91.5 persen pada parameter C sebesar 0.01. Hasil optimasi ini akan digunakan untuk langkah analisis selanjutnya yaitu menyusun *confusion matrix*. *Confusion matrix* adalah matriks yang menunjukkan kinerja algoritma dalam mengklasifikasi secara visual. Melalui matriks ini dapat diketahui perbandingan antara klasifikasi aktual dengan prediksinya. Berikut merupakan hasil *confusion matrix* untuk kernel linier dengan parameter C sebesar 0.01.

Tabel 4 merupakan*confusion matrix* yang didapat menggunakan data testing yaitu 25 persen dari dataset. Berdasarkan Tabel 3 dapat diketahui jika status ketertinggalan wilayah terbagi menjadi status tertinggal dan status tidak tertinggal.

Tabel 4. *Confusion Matrik* Kernel Linier

	Aktual	
Prediksi	Tertinggal	Tidak Tertinggal
Tertinggal	103 ***True Positive (TP)***	11 ***False Positive (FP)***
Tidak Tertinggal	0 ***False Negative (FN)***	15 ***True Negative (TN)***

Terdapat 129 data yang termasuk kedalam data testing yang digunakan untuk menyusun *confusion matrix*. Berdasarkan *confusion matrix* dapat diketahui, SVM dengan kernel linier memprediksi status wilayah "Tertinggal" adalah sebanyak 114 kali, dan memprediksi suatu wilayah kedalam status "Tidak Tertinggal" adalah sebanyak 15 kali. Sementara itu 103 daerah secara data aktual merupakan daerah berstatus tertinggal dan 26 daerah lainnya adalah daerah berstatus tidak tertinggal. Berdasarkan Tabel 4 nilai prediksi bernilai 11 merupakan nilai *False Positive* (FP), artinya nilai tersebut adalah prediksi yang menetapkan suatu daerah berstatus tertinggal sedangkan secara data aktualnya daerah tersebut tidak berstatus tertinggal. Selanjutnya nilai prediksi sebesar 0 merupakan nilai *False Negative* (FN), artinya nilai tersebut adalah prediksi yang menetapkansuatu daerah berstatus tidak tertinggal sedangkan secara aktualnya daerah berstatus tertinggal. Berdasarkan hasil *confusion matrix*, nilai-nilai yang terdapat didalam *confusion matrix* pada Tabel 4 dapat digunakan untuk menghitung nilai akurasi, nilai ini yang menujukkan performa dari kernel linier dalam mengklasifikasikan status ketertinggalan wilayah. Berikut merupakan hasil perhitungan akurasi menggunakan kernel linier.

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} = \frac{103+15}{103+15+0+11} = 0{,}915 \quad (1)$$

Hasil perhitungan akurasi pada persamaan (1) menunjukkan bahwa performa klasifikasi menggunakan algoritma SVM dengan kernel linier sebesar 0.915 atau setara dengan 91.5% yaitu mampu dengan tepat mengklasifikasikan 118 sampel dari total 129 sampel data testing.

Klasifikasi Daerah Tertinggal Menggunakan Radial Basis Function (RBF) Kernel SVM

Klasifikasi kedua yaitu klasifikasi menggunakan kernel RBF. Berbeda dengan kernel linier, kernel RBF adalah salah satu

fungsi dalam SVM yang digunakan untuk karakteristik data yang tidak terindikasi terklasifikasi secara linier. Kernel RBF pada SVM memiliki parameter C atau *Cost*dan parameter Gamma (γ). Kernel RBF bekerja dengan cara mengoptimasi parameter C dan Gammauntuk mendapatkan akurasi klasifikasi terbaik dengan cara melakukan berbagai kombinasi model atau *trial and error*.Hasil optimasi parameter C dan γ menggunakan fungsi kernel RBF adalah sebagai berikut.

Tabel 5. Hasil Optimasi Pemilihan Parameter Terbaik Kernel RBF

C	Gamma(γ)				
	0.01	0.02	0.03	0.04	0.05
10^{-3}	0.798	0.798	0.798	0.798	0.798
10^{-2}	0.798	0.798	0.798	0.798	0.798
10^{-1}	0.837	0.876	0.884	0.868	0.860
1	0.915	0.907	0.922	0.922	**0.922***

Ket : *) Parameter terpilih dengan nilai akurasi terbesar

Tabel 5 menunjukkan hasil optimasi parameter C dan γmenggunakan kernel RBF, hasil tersebut didapat menggunakan data training yaitu 75 persen dari dataset. Berdasarkan hasil optimasi tersebut didapatkan nilai akurasi terbaik yaitu 0.922 atau 92.2 persen pada parameter C = 1, serta γ = 0.05. Hasil optimasi ini akan digunakan untuk langkah analisis selanjutnya yaitu menyusun *confusion matrix.* Berikut merupakan hasil *confusion matrix* untuk kernel RBF dengan parameter C sebesar 1 serta γ = 0.05.

Tabel 6. *Confusion Matrix* Kernel RBF

	Aktual	
Prediksi	Tertinggal	Tidak Tertinggal
Tertinggal	101 ***True Positive (TP)***	8 ***False Positive (FP)***
Tidak Tertinggal	2 ***False Negative (FN)***	18 ***True Negative (TN)***

Tabel 6 merupakan *confusion matrix* yang didapat menggunakan data testing yaitu 25 persen dari dataset. Berdasarkan Tabel 6 dapat diketahui jika status ketertinggalan wilayah terbagi menjadi status tertinggal dan status tidak tertinggal. Terdapat 129 data yang termasuk kedalam data testing yang digunakan untuk menyusun *confusion matrix*. Berdasarkan *confusion matrix* dapat diketahui, SVM dengankernel RBF memprediksi status wilayah "Tertinggal" adalah sebanyak 109 kali, dan memprediksi suatu wilayah kedalam status "Tidak Tertinggal" adalah sebanyak 20 kali. Sementara itu 103 daerah secara data aktual merupakan daerah berstatus tertinggal dan 26 daerah lainnya adalah daerah berstatus tidak tertinggal.Berdasarkan Tabel 6 nilai prediksi bernilai 8 merupakan nilai *False Positive* (FP), artinya nilai tersebut adalah prediksi yang menetapkan suatu daerah berstatus tertinggal sedangkan secara data aktualnya daerah tersebut tidak berstatus tertinggal. Selanjutnya nilai prediksi sebesar 2 merupakan nilai *False Negative* (FN), artinya nilai tersebut adalah prediksi yang menetapkansuatu daerah berstatus tidak tertinggal sedangkan secara aktualnya daerah berstatus tertinggal. Berdasarkan hasil *confusion matrix*, nilai-nilai yang terdapat didalam *confusion matrix* pada Tabel 6 dapat digunakan untuk menghitung nilai akurasi, nilai ini yang menujukkan performa dari kernel RBF dalam mengklasifikasikan status ketertinggalan. Berikut merupakan hasil perhitungan akurasi menggunakan kernel RBF

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} = \frac{101+18}{101+18+2+8} = 0{,}922 \quad (2)$$

Hasil perhitungan akurasi pada persamaan (2) menunjukkan bahwa performa klasifikasi menggunakan algoritma SVM dengan kernel RBF sebesar 0.922 atau setara dengan 92.2% yaitu mampu dengan tepat mengklasifikasikan 119 sampel dari total 129 sampel data testing.

Klasifikasi Daerah Tertinggal Menggunakan Polinomial Kernel SVM

Klasifikasi terakhir untuk algoritma SVM adalah fungsi kernel polinomial. Setiap data pengamatan tentunya memiliki perbedaan karaktersitik, salah satunya adalah data dengan karakteristik non linier. Kernel polinomial merupakan fungsi kernel yang memfasiltasi untuk jenis data yang non linier. Kernel ini pada SVM memiliki parameter *Cost* (C) dan *Degree* (d). Penentuan parameter terbaik SVM menggunakan polinomial kernel dapat dievaluasi melalui ukuran performa klasifikasi yang dapat diukur menggunakan akurasi. Hasil optimasi parameter C dan dmenggunakan fungsi kernel polinomial adalah sebagai berikut.

Tabel 7 menunjukkan hasil optimasi parameter C dan d menggunakan kernel polinomial. Hasil tersebut didapat

menggunakan data training yaitu 75 persen dari dataset. Berdasarkan hasil optimasi tersebut didapatkan nilai akurasi terbaik yaitu 0.907 atau 90.7 persen pada parameter C = 1, serta d = 1.

Tabel 7. Hasil Optimasi Pemilihan Parameter Terbaik Kernel Polinomial

	Degree(d)				
C	1	2	3	4	5
10^{-3}	0.798	0.798	0.806	0.806	0.822
10^{-2}	0.806	0.806	0.837	0.837	0.837
10^{-1}	0.899	0.860	0.868	0.853	0.853
1	**0.907**	0.884	0.884	0.868	0.868

Ket : *) Parameter terpilih dengan nilai akurasi terbesar

Hasil optimasi ini digunakan untuk langkah analisis selanjutnya yaitu menyusun *confusion matrix.* Berikut merupakan hasil *confusion matrix* untuk kernel polinomial dengan parameter C sebesar 1 serta d = 1.

Tabel 8. *Confusion Matrix* Kernel Polinomial

	Aktual	
Prediksi	Tertinggal	Tidak Tertinggal
Tertinggal	101 ***True Positive (TP)***	10 ***False Positive (FP)***
Tidak Tertinggal	2 ***False Negative (FN)***	16 ***True Negative (TN)***

Tabel 8 merupakan*confusion matrix* yang didapat menggunakan data testing yaitu 25 persen dari dataset. Berdasarkan Tabel 8 dapat diketahui jika status ketertinggalan wilayah terbagi menjadi status tertinggal dan status tidak tertinggal. Terdapat 129 data yang termasuk kedalam data testing yang digunakan untuk menyusun *confusion matrix*. Berdasarkan *confusion matrix* dapat diketahui, SVM dengankernel polinomial memprediksi status wilayah "Tertinggal" adalah sebanyak 111 kali, dan memprediksi suatu wilayah kedalam status "Tidak Tertinggal" adalah sebanyak 18 kali. Sementara itu 103 daerah secara data aktual merupakan daerah berstatus tertinggal dan 26 daerah lainnya adalah daerah berstatus tidak tertinggal.Berdasarkan Tabel 8 nilai prediksi bernilai 10 merupakan nilai *False Positive* (FP), artinya nilai tersebut adalah prediksi yang menetapkan suatu daerah berstatus tertinggal sedangkan secara data aktualnya daerah tersebut tidak berstatus tertinggal. Selanjutnya nilai prediksi sebesar 2 merupakan nilai *False Negative* (FN), artinya nilai tersebut adalah prediksi yang menetapkan suatu daerah berstatus tidak tertinggal sedangkan secara aktualnya daerah berstatus tertinggal. Berdasarkan hasil *confusion matrix*, nilai-nilai yang terdapat didalam *confusion matrix* pada Tabel 8 dapat digunakan untuk menghitung nilai akurasi, nilai ini yang menujukkan performa dari kernel polinomial dalam mengklasifikasikan status ketertinggalan. Berikut merupakan hasil perhitungan akurasi menggunakan kernel polynomial.

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} = \frac{101+16}{101+16+2+10} = 0{,}907 \quad (3)$$

Hasil perhitungan akurasi pada persamaan (3) menunjukkan bahwa performa klasifikasi menggunakan algoritma SVM dengan kernel polinomial sebesar 0.907 atau setara dengan 90.7% yaitu mampu dengan tepat mengklasifikasikan 117 sampel dari total 129 sampel data testing.

Klasifikasi Daerah Tertinggal Menggunakan *k*-*Nearset Neighbor* (*k*-NN)

Selain menggunakan algoritma SVM, pada penelitian ini juga menggunakan algoritma *k-Nearset Neighbor* (*k*-NN) yang merupakan jenis dari *supervised machine learning* yang sistem kerja klasifikasinya memperhitingkan *distance* atau jarak antar data pengamatan. Sama halnya dengan SVM yang memiliki parameter untuk melakukan klasifikasi, *k*-NN juga memiliki parameter yang disimbolkan dengan *k* yang nantinya dilakukan optimasi untuk menentukan parameter terbaik dengan hasil performa klasifikasi terbaik. Berikut merupakan hasil optimasi parameter k pada algoritma *k*-NN untuk mengklasifikasikan status ketertinggalan daerah di Indonesia.

Tabel 9 menunjukkan hasil optimasi parameter k menggunakan algoritma k-NN. Hasil tersebut didapat menggunakan data training yaitu 75 persen dari dataset. Berdasarkan hasil optimasi tersebut didapatkan nilai akurasi terbaik yaitu 0.922 atau 92.2 persen pada parameter *k* = 15. Hasil optimasi ini akan digunakan untuk langkah analisis selanjutnya yaitu menyusun *confusion matrix.* Berikut merupakan hasil *confusion matrix* untuk algoritma *k*-NN dengan parameter k sebesar 15.

Tabel 9. Hasil Optimasi Pemilihan Parameter Terbaik k Nearset Neighbor (*k*-NN)

Parameter	Akurasi	Parameter	Akurasi
k=1	0.822	k=9	0.892
k=2	0.868	k=10	0.892
k=3	0.876	k=11	0.899
k=4	0.899	k=12	0.899
k=5	0.899	k=13	0.899
k=6	0.899	k=14	0.906
k=7	0.892	**k=15**	**0.922**
k=8	0.899		

Tabel 10. *Confusion Matrix k*-NN

	Aktual	
Prediksi	Tertinggal	Tidak Tertinggal
Tertinggal	102 ***True Positive (TP)***	9 ***False Positive (FP)***
Tidak Tertinggal	1 ***False Negative (FN)***	17 ***True Negative (TN)***

Tabel 10 merupakan *confusion matrix* yang didapat menggunakan data testing yaitu 25 persen dari dataset. Berdasarkan Tabel 10 dapat diketahui jika status ketertinggalan wilayah terbagi menjadi status tertinggal dan status tidak tertinggal. Terdapat 129 data yang termasuk kedalam data testing yang digunakan untuk menyusun *confusion matrix*.

Berdasarkan *confusion matrix* dapat diketahui, *k*-NN dengan parameter *k* sebesar 15 memprediksi status wilayah "Tertinggal" adalah sebanyak 111 kali, dan memprediksi suatu wilayah kedalam status "Tidak Tertinggal" adalah sebanyak 18 kali.

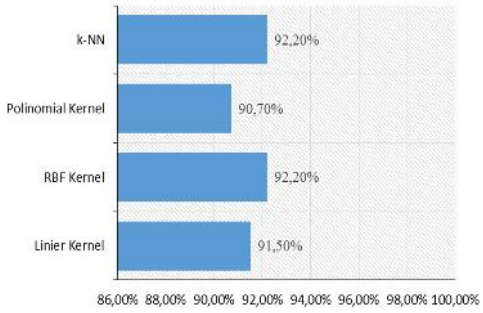
Sementara itu 103 daerah secara data aktual merupakan daerah berstatus tertinggal dan 26 daerah lainnya adalah daerah berstatus tidak tertinggal. Berdasarkan Tabel 10 nilai prediksi bernilai 9 merupakan nilai *False Positive* (FP), artinya nilai tersebut adalah prediksi yang menetapkan suatu daerah berstatus tertinggal sedangkan secara data aktualnya daerah tersebut tidak berstatus tertinggal. Selanjutnya nilai prediksi sebesar 1 merupakan nilai *False Negative* (FN), artinya nilai tersebut adalah prediksi yang menetapkan suatu daerah berstatus tidak tertinggal sedangkan secara aktualnya daerah berstatus tertinggal. Berdasarkan hasil *confusion matrix*, nilai-nilai yang terdapat didalam *confusion matrix* pada Tabel 10 dapat digunakan untuk menghitung nilai akurasi, nilai ini yang menujukkan performa dari *k*-NN dalam mengklasifikasikan status ketertinggalan. Berikut merupakan hasil perhitungan akurasi menggunakan *k*-NN.

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} = \frac{102+17}{102+17+1+9} = 0{,}922 \quad (4)$$

Hasil perhitungan akurasi pada persamaan (4) menunjukkan bahwa performa klasifikasi menggunakan algoritma *k*-NN sebesar 0.922 atau setara dengan 92.2% yaitu mampu dengan tepat mengklasifikasikan 119 sampel dari total 129 sampel data testing.

Perbandingan Klasifikasi Daerah Tertinggal Menggunakan *k*-NN dan SVM

Tujuan dari penelitian ini adalah untuk mengetahui hasil ketepatan yang paling baik diantara algoritma SVM dan *k*-NN. Oleh karena itu tahap selanjutnya adalah membandingan diantara kedua algoritma tersebut berdasarkan performa klasifikasinya. Berikut merupakan visualisasi hasil perbandingan performa klasifikasi masing-masing algoritma.



Gambar 2. Perbandingan Nilai Akurasi k-NN dan SVM

Performa klasifikasi yaitu nilai akurasi dikategorikan berdasarkan beberapa kelompok yang disajikan pada Tabel 11.

Tabel 11. Klasifikasi Tingkat Akurasi (Aulianita, 2016)

Akurasi(%)	Performa
>90 – 100	Sangat Baik
>80 – ≤ 90	Baik
>70 – ≤ 80	Cukup
60 – ≤ 70	Buruk
< 60	Salah

Berdasarkan Tabel 11 dapat disimpulkan bahwa hasil klasifikasi daerah tertinggal dengan algoritma *Support Vector Machine* (SVM) untuk fungsi kernel RBF dan algoritma *k*-NN merupakan algoritma dengan kualifikasi

sangat baik. Kedua algoritma tersebut jika dibandingkan memiliki nilai akurasi yang sama baik dan dapat digunakan untuk menentukan klasifikasi daerah tertinggal.

KESIMPULAN

Berdasarkan hasil klasifikasi daerah tertinggal dengan algoritma SVM fungsi RBF kernel memiliki hasil terbaik dengan parameter C=1 serta γ=0.05 yang memiliki performa klafiskasi sebesar 92.2%. Sedangkan hasil klasifikasi daerah tertinggal dengan algoritmak-NN diperoleh hasil terbaik dengan *k*=15 yang memiliki performa klafiskasi sebesar 92.2%. Berdasarkan hasil klasifikasi daerah tertinggal dengan algoritmaa SVM dan *k*-NN termasuk dalam performa sangat baik. Kedua metode tersebut jika dibandingkan memiliki nilai akurasi yang sama baik dan dapat digunakan untuk menentukan klasifikasi daerah tertinggal.

DAFTAR PUSTAKA

Abe S. 2010. *Support Vector Machines for Pattern Classification 2nd Edition.* London: Springer-Verlag.

Al Azies H. 2017. *Analisis Perilaku Hidup Bersih Dan Sehat (PHBS) Rumah Tangga Penderita TB Di Wilayah Pesisir Kota Surabaya Menggunakan Pendekatan Regresi Logistik Biner*. [Skripsi, Institut Teknologi Sepuluh Nopember]

Al Azies H, Trishnanti D, Mustikawati EPH. 2019.Comparison of Kernel Support Vector Machine (SVM) in Classification of Human Development Index (HDI), *IPTEK Journal of Proceedings Series*. 1:53-57.

Aulianita, Rizki. 2016. Komparasi Metode K-Nearest Neighbors dan Support Vector Machine Pada Sentiment Analysis Review Kamera. *Journal Speed-Sentra Penelitian Engineering dan Edukasi.* 8(3):71-77.

Ayodele, TO. 2010. *New Advances in Machine Learning*, Yagang Zhang (Ed). London: IntechOpen Limited.

Delgado M, Cernadas E, Barro, S, & AmorimD. 2014. Do we need hundreds of classifiers to solve real world classification problems?.*The Journal of Machine Learning Research.* 15:3133-3181.

Deng Z, Zhu X, Cheng D, Zong M, Zhang S. 2016. Efficient *k*-NN classification algorithm for big data. *Neurocomputing.* 195: 143–148.

Fernanda, J W, Anuraga G, Fahmi, MA. 2019. Risk factor analysis of hypertension with logistic regression and Classification and Regression Tree (CART). In *Journal of Physics: Conference Series*. 1217(1): 012109.

Gunn S. 1998. *Support Vector Machine for Clasification and Regression*. Southamton: University of Southampton Institutional Repository.

Guo G., Wang H., Bell D., Bi Y., Greer K. 2003 KNN Model-Based Approach in Classification. *In: Meersman R., Tari Z., Schmidt D.C. (eds) On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM 2003. Lecture Notes in Computer Science*. 2888: 986-996.

James G, Witten D, Hastie T, Tibshirani R. 2013. *An introduction to statistical learning: with applications in R.* New York:Springer.

Jung M, Niculita O, Skaf Z. 2018. Comparison of different classification algorithms for fault detection and fault isolation in complex systems. *Procedia Manufacturing.*19:111-118.

Kotsiantis SB. 2007. Supervised Machine Learning: A Review of Classification Techniques. *Informatica.* 31:249-268.

Purwandari T, Hidayat Y. 2017. Pemodelan Ketertinggalan Daerah di Indonesia Menggunakan Analisis Diskriminan**.** *Prosiding Konferensi Nasional Penelitian Matematika dan Pembelajarannya (KNPMP)*. 2: 194-200.

Puspitasari D A, Rustam Z. 2018. Application of SVM-KNN using SVR as feature selection on stock analysis for Indonesia stock exchange. In *AIP Conference Proceedings*. 2023:020207.

Russel, S. J. dan Norvig, P. (2016), Artificial intelligence: a modern approach, Malaysia; Pearson Education Limited

Shalev-Shwartz S, Ben-David S. (2014). *Understanding Machine Learning From Theory to Algorithms.*UK: Cambridge University Press.

Smola A, Vishwanathan SVN. 2008. *Introduction to machine learning*. UK: Cambridge University Press.

Tan PN, Steinbach M, Karpatne A, Kumar V. 2019. *Introduction to Data Mining, 2nd Edition*. London: Pearson Education, Inc.

Vapnik VN. 1995. *The Nature of Statistical Learning Theory (2nd ed.)*. Springer Verlag.